AUTHOR IDENTIFICATION IN NOVELS

Chirag Gupta,Harshit Maheshwari
Advisor : Professor Ajai Jain
Dept of Computer Science and Engineering
IIT Kanpur - 208 016
INDIA

16 July 2012

INDEX

ACKNOWLEDGEMENT

Firstly, we express our thanks to Professor Ajai Jain who gave us this opportunity to learn the subject with a practical approach, guided us and gave us valuable suggestions regarding the project report.

Through this acknowledgment, we express our sincere gratitude to all those people who have been associated with this assignment and have helped us with it and made it a worthwhile experience.

Finally we extend our thanks to the various people who have shared their opinions and experiences through which we received the required information crucial for our report.

ABSTRACT

Author Identification is an area with a wide scope of further research. Different authors have different styles of writing. Its these small style differences that separate one author from another.

In the past, research has been performed on author identification of e-mails by O. De Vel,A.Anderson,M. Corney and G. Mohay. Their datasets included 3 authors' emails over 3 topics or categories and they used nearly 10 emails for 3 authors per categories. They applied Support Vector Machine (SVM) algorithm on features such as N-graphs(where N=2). They could give good categorisation for 2 of those 3 authors.

Author identification in short texts has been done by Marcia Fissette who on the basis of features like unigrams,bigrams,triplets and smileys could distinguish on a set of short texts of 40 authors with accuracy varying from 5-12 % . Also on Author identification in short texts C. Chaski has had results of 95% accuracy using syntactic analysis of texts.

Our project differs from the above as the above are author identification for short texts and we wish to identify authors in a dataset of novels.

This can be applied in areas like: detecting text forgeries in universities where students have to write essays as assignments. Then copied text can be found out easily.

Also, the data on wikipedia pages can be edited by indivuals across the world. Most of these edits are spams and have to be checked manually before making the final changes. If we can ascertain the theme of the edits and check them against the theme of the original article then we can safely reduce a lot of manual work.

English literature comprises of authors of wide varieties varying in different eras, genres and complexities. It is a need to be able to narrow down an unknown novel to a single or couple of authors so as to reduce chances of forgery.

The question arises how to capture this difference in writing styles merely by looking at the texts?

INTRODUCTION

In this project, we have tried to capture this difference in styles between authors using different techniques; techniques specific to each of the above varying features.

Speaking in layman's language: First of all we analyse the writing features of the input, compare them with those of the authors we already have in our dataset, multiply values pertaining to features which are useful (selected on the basis of graphical analysis) and output the two authors with the highest values.

We did our work on a dataset of 11 authors which are :

- Agatha Christie
- Enid Blyton
- Charles Dickens
- Ian Fleming
- Jane Austen
- Jeffery Archer
- Mark Twain
- Oscar Wilde
- P.G.Wodehouse
- Robert Ludlum
- Rudyard Kipling

We used a total number of 198 novels of the above authors in .txt format to form our dataset and 42 novels to apply our tests.

OUR APPROACH

These were the features that we found to be really heplful in narrowing down our search:

- Era
- Genre
- Sentence Length
- Sentence Depth
- Punctuations used(. , : - ? !)

In this project we tried to capture the different writing styles of the authors by comparing them on different parameters. The parameters used were:

- **Era**: First of all we try to classify the authors into different eras based on their vocabulary. We compliled a list of frequently used words used by the authors of 16th century, 18th century and 20th century. There were some words that occurred in 2 categories and some authors belonged to both the 18th and the 20th century because there is no strict timeline difference between the 2 centuries. We counted the number of matching words of the authors from each era and divided them by the total number of words of the particular era. The maximum relative frequency of the words used among all the eras classified the era of the author.
- **Genres**: We then classify the authors based on their genres. We compiled a list of frequently used words for the following genres:
  - Crime
  - Detective/Mystery
  - Fantasy
  - Horror
  - Religion
  - Science fiction
  - Romance

Based on the word list we classified the novel into one of the above genres. We counted the number of matching words of the authors from each genre and divided it by total number of words of the particular genre. The maximum relative frequency of the words used classified the genre of the particular novel.

Also, if the relative frequency of a particular genre came quite close to the maximum relative frequency of a genre then we classified the novel in that genre too. Each author usually sticks to a few of the above mentioned genres and this further helps in differntiation of the authors.

- **Sentence Length**: Different authors use different sentence lengths in their writing styles. We define sentence length as the number of words in each sentence. A sentence is terminated by either a '.', '!' or a '?' . We found that the average sentence length of the authors (defined as the Σ sentence lengths / number of sentences) for each author is usually consistent within his different works as depicted by the graph 1.1. Using this parameter we could differntiate between the authors to some extent.

Sentence Lengths

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
8.96	17.04	10.33	11.34	11.54	13.99	16.02	16.34	10.11	10.67	13.74
10.39	20.96	10.2	10.33	13.68	12.84	13.94	9	12.08	11.58	12.43
8.79	28.88	10.55	11.96	13.1	15.24	15.04	8.49	11.93	10.94	15.14
7.85	20.09	10.57	12.1	15.16	12.8	14.63	15.36	10.94	10.3	10.55
6.66	14.68	10.11	12.76	15.67		16.31	8.27	10.68	11.48	13.51
7.67	17.51	9.04	11.49	17.42		14.49	11.68	9.83	10.03	4.18
7.42	18.12	10.57	13.19	20.28		16.55	14.58	8.63	11.44	14.39
8.33	17.09	10.11	13.19			18.61	9.39	11.32	10.12	11.89
8.38	16.86	10.88				22.11		14.33	9.04	10.6
7.77	13.82	11.8				14.25			12.14	
7.4	14.86	8.47				15.11			12.43	
8.34	11.74	8.84				16.65			9.81	
8.74	12.99	9.07				17.8			10.95	
7.53	18.16	9.49				16.02				
8.5	17.16					20.12				
7.45	15.92					14.96				
9.52	24.93					18				
7.53	25.28					19.63				
8.84	21.55					20.89				
8.87	21.42					19.05				
7.67	14.34					17.8				
8.96	12.84					14.94				
7.87	14.66					16.02				
10.39	16.97					17.83				
	30.92					17.3				
	19.75					17.67				
	18.82					17				
	19.54					17.23				
	30.34					19.29				
	17.05					17.69				
	33.95					16.11				
	33.57					14.18				
	14.05					15.96				
	12.96					18.15				
	15.12					17.08				
	21.33					19.04				
	15.74					15.75				
	14.25					16.78				
	19.31									
	23.79									
	23.41									
	20.57									
	13.27									
	16.36									

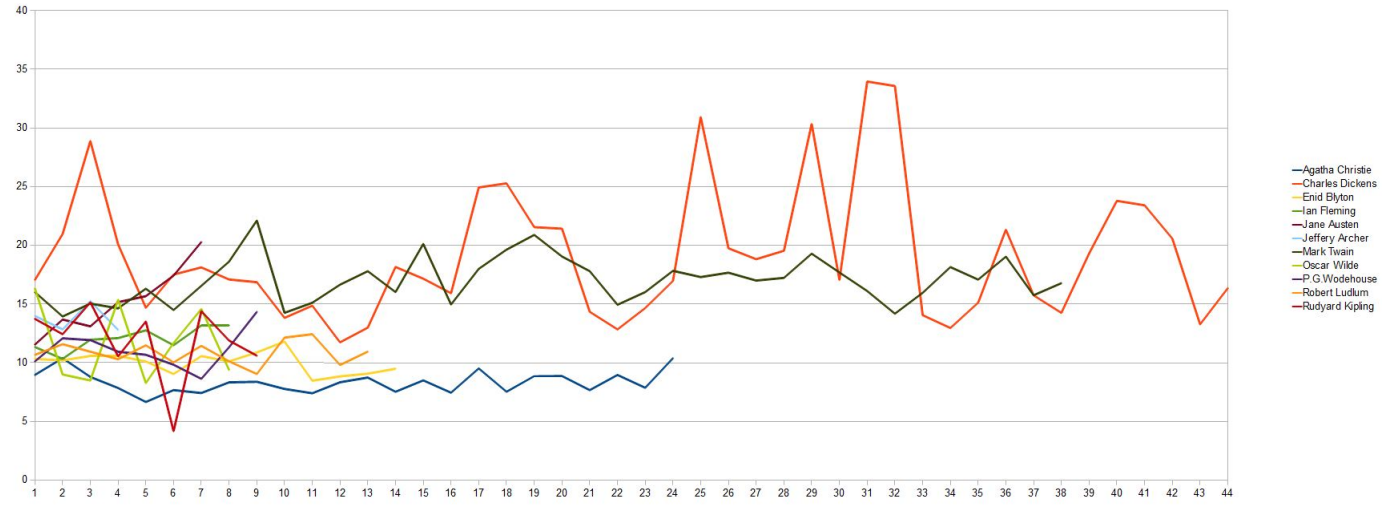

Figure 1.1 : Sentence Lengths

- **Sentence Depth** : This parameter to some extent sheds light on the semantic style of the authors. We parse the sentences using the stanford parser and look at the parsed tree thus formed. The sentence depth is defined as the height of the parsed tree and authors using more complex sentences have more depth than authors using simple sentence structure. An example parsed tree is shown in the figure below.

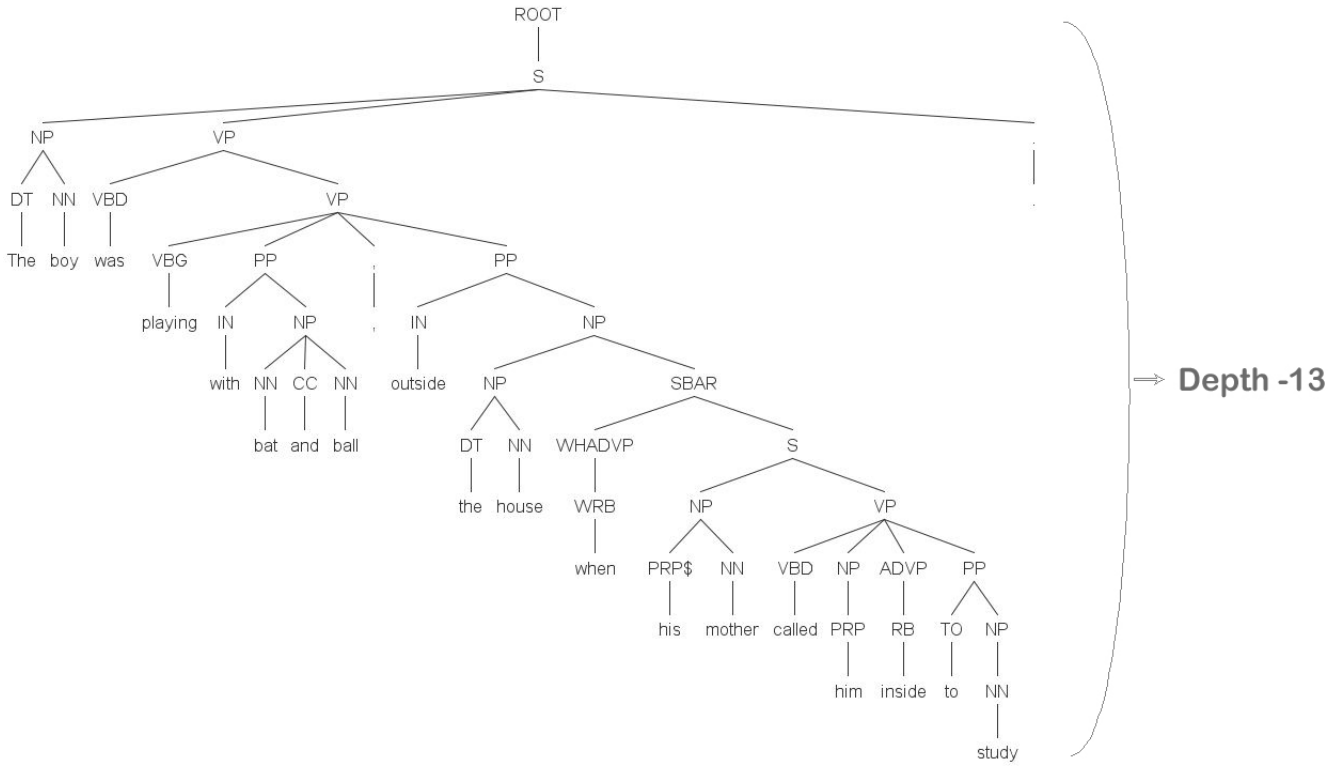

Figure : Sentence Tree

This depth we found using Stanford's parser. It parsed each sentence into a tree giving a sequence of brackets and tokens for the sentence as output. In this output, the maximum number of open brackets "(" at any point of time is a sentence's depth. We calculated depth for each sentence and averaged it over a novel. This thus gives an idea about the complexity of a sentence.

We used the average sentence depth of the novel (defined as the Σ sentence depth/total sentences) for each author and found that the average sentence depth is also consistent within his different works as depicted by the graph 1.2 . Authors like Enid Blyton who usually write for small children has small average sentence depth as in comparison with writers like Mark Twain. This further seperates the authors to some extent.

Sentence Depths

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
8.52	10.08	8.29	8.08	9.31	10.28	10.43	9.82	8.15	8.38	9.28
7.63	11.3	8.26	8.23	9.92	9.86	9.58	8.18	9.43	8.46	8.67
8.37	11.28	8.18	8.5	9.93	10.58	9.75	7.94	9.18	8.4	10.05
8.09	9.81	8.21	8.62	10.14	10.06	10.02	10.18	8.58	8.67	9.03
8.03	9.49	8.07	8.72	10.61		9.95	7.72	8.97	8.33	9.59
8.33	10.04	8.25	8.47	10.68		9.97	9.29	8.74	7.89	7.49
8.13	10.2	8.21	8.9	11.11		10.38	9.7	8.42	9.1	9.76
8.64	10.35	8.07	8.71			10.9	8.28	9.2	7.4	8.76
8.19	10.24	8.86				11.22		9.55	8.09	9.19
8.39	9.62	8.95				9.51			8.27	
8.27	9.08	8.02				10.06			8.12	
8.16	9.03	8.32				10.68			8.41	
8.33	9.61	8.4				10.64			8.85	
8.04	10.24	8.58				10.42				
8.79	10.24					11.14				
8.21	9.78					9.69				
8.51	10.98					10.66				
8.04	11.6					11.65				
9.02	10.27					10.53				
8.16	11.5					11.26				
8.02	9.32					10.54				
8.52	9.34					10.11				
8.02	9.76					10.43				
7.63	9.99					10.18				
	11.01					10.33				
	10.23					11.21				
	10.33					10.53				
	10.65					10.53				
	11.68					10.63				
	9.49					10.57				
	12.95					10.24				
	12.38					9.39				
	8.72					10.37				
	8.72					10.4				
	8.77					10.42				
	10.65					10.95				
	9.52					9.97				
	9.41					11.13				
	10.19									
	11.25									
	11.67									
	10.63									
	8.96									
	10.13									

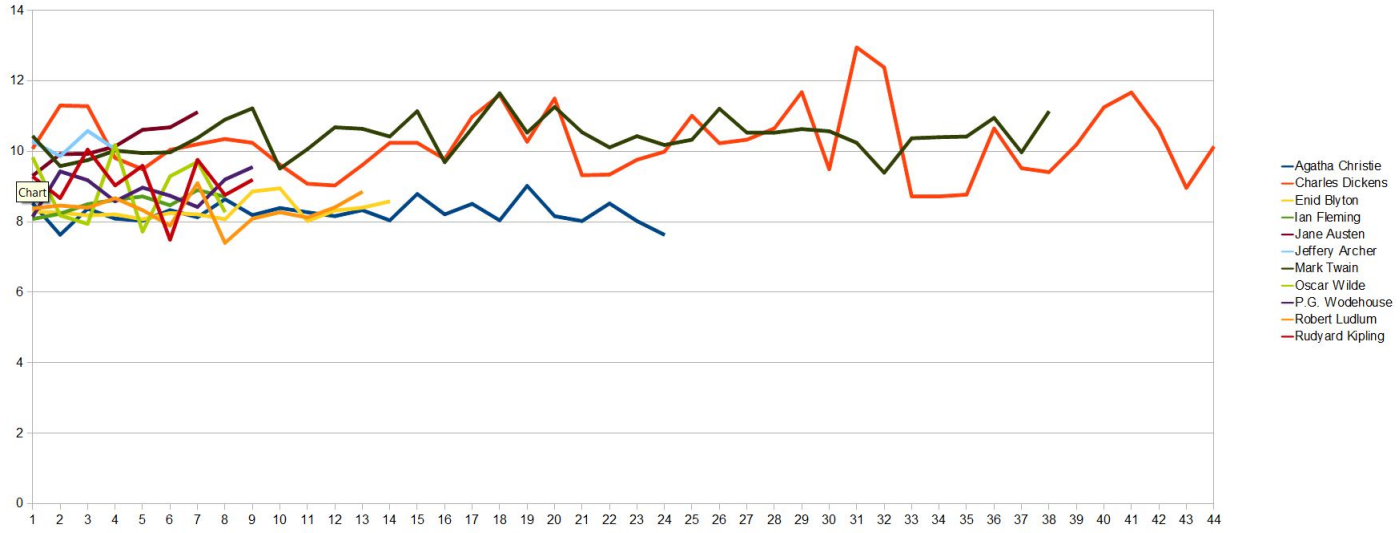

Figure 1.2 : Sentence Depths

- **Punctuations** : The authors also have a pattern for punctuation usage. We took the following punctuations and analyzed them for the authors:
  - Comma ’,’
  - Dot ’.’
  - Question mark ’?’
  - Semi Colon ’;’
  - Colon ’:’
  - Quotes ’’
  - Exclamation mark ’!’
  - Hyphen ’-’

We found some punctuations to be useful parameters while some of the punctuations were completely useless as depicted in the graphs below.
We also found that some authors used a very consistent number of relative punctuations like :
Jane Austen used a nearly the same relative freqeuncy of hyphens in all her novels as depicted in the graph 1.7. For this reason if the relative frequency of a test novel came quite close to that relative frequency (of Jane Austen) we increase the probability of the novel’s author being Jane Austen.
Similarly the relative frequency of the number of colons used by Enid Blyton was quite consistent in all her novels. So if a new test case novel had the same relative number of colons then we increased the probability of the author being Enid Blyton.

Colon(:)

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
0.41	0.16	0.01	0.6	0.19	0.12	0.8	0.21	0.11	0.12	0.24
0.35	0.38	0.01	0.15	0.34	0.13	0.62	0.21	0	0.07	0.22
0.28	0.91	0.01	0.11	0.21	0.24	0.59	0.08	0.21	0.19	0.14
1.27	0.23	0.01	0.65	0.07	0.16	0.62	0.23	0.11	0.11	0.36
2.96	0.03	0	0.05	0.08		0.1	0.05	0.09	0.21	0.22
1.76	0.21	0.01	0.06	0.19		0.92	1.35	0.07	0.14	0.14
2.89	0.35	0.01	0.15	0.28		0.57	0.06	0.06	0.33	0.1
1	0.21	0	0.16			0.28	0.01	0.68	0.24	0.22
0.2	0.47	0				0.55		0	0.24	0.07
0.31	0.25	0.01				1.05			0.05	
0.52	0.11	0.01				0.18			0.12	
0.08	0.11	0.03				0.28			0.32	
0.06	0.09	0				0.33			0.36	
0.26	0.21	0.01				0.33				
0.19	0.18					0.59				
0.7	0.15					0.31				
0.29	0.09					0.55				
0.26	0.25					0.51				
0.07	0.08					0.24				
0.07	0.1					0.69				
0.86	0.48					0.25				
0.41	0.33					1.17				
0.01	0.31					0.82				
0.35	0.33					0.24				
	1.02					0.47				
	0.2					0.5				
	0.27					0.77				
	0.2					0.74				
	0.21					0.32				
	0.23					0.47				
	0.37					0.35				
	0.48					1.02				
	0.3					0.75				
	0.5					0.9				
	0.25					0.98				
	0.09					0.39				
	0.23					0.48				
	0.55					0.62				
	0.09									
	0.26									
	0.18									
	0.11									
	0.32									
	0.29									

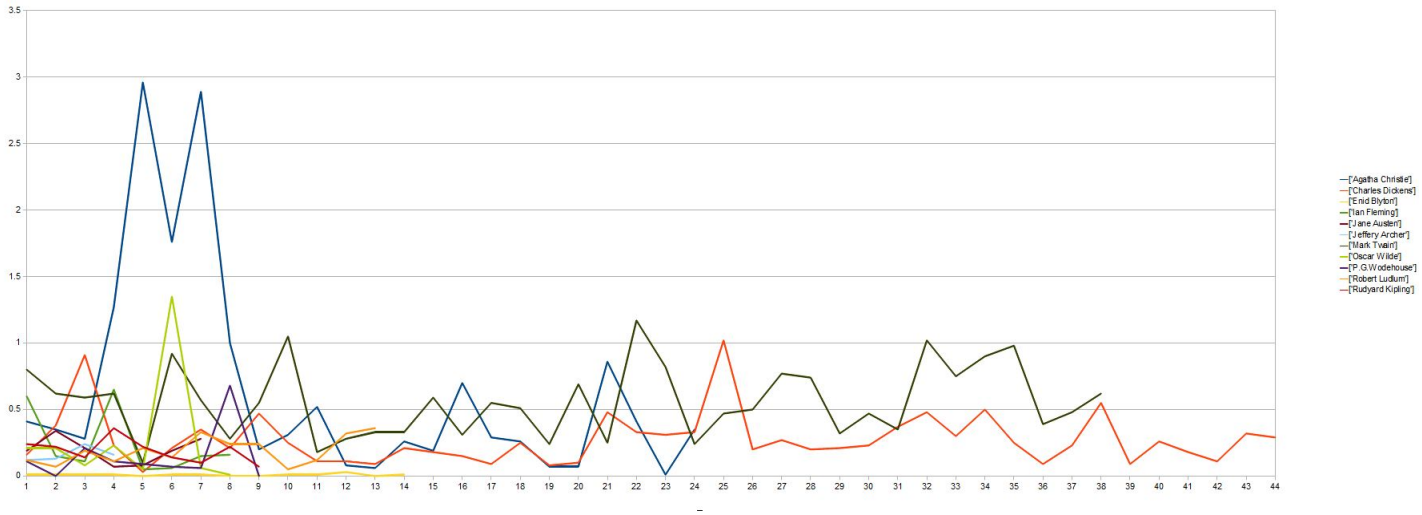

Figure 1.3 : Colon

Comma(,)

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
13.25	18.39	13.2	7.54	13.44	9.5	13.7	15.31	15.69	11.02	13.92
9.97	14.77	12.54	9.63	13.49	10.28	11.96	13.34	12.21	11.13	14.64
10.85	16.07	12.3	7.81	13.24	8.87	13.42	11.75	12.82	9.02	15.27
9.73	18.03	10.9	6.98	10.58	10.19	12.74	12.23	13.28	11.07	11.84
10.07	14.95	11.83	10.19	14.47		14.63	11.92	10.45	9.36	13.66
10.65	15.73	12.35	10.99	14		12.7	16.66	12.58	9.1	9.08
8.65	19.04	10.9	11.32	15.09		10.58	15	13.73	12.28	14
12.15	15.34	11.83	9.38			9.88	11.98	15.06	8.34	14.39
10.9	17.15	12.69				15.97		11.44	9.16	11.95
11.38	18.07	12.8				12.9			14.39	
10.93	20.63	13.65				11.06			13.21	
11.29	17.25	12.4				8.63			11.99	
11.44	17.11	11.79				11.34			11.87	
12.3	17.13	12.87				12.17				
10.79	16					10.47				
11.5	17.2					15.94				
11.73	15.38					13.68				
12.33	14.92					9.62				
13.54	7.52					11.23				
11.72	14.14					11.37				
10.86	16.93					12.38				
13.25	14.98					11.5				
11.01	17.29					14.04				
9.97	18.95					14.35				
	19.03					12.82				
	17.18					11.31				
	16.23					10.63				
	15.89					12.12				
	15.77					11.83				
	16.98					12.17				
	15.04					12.95				
	15.22					10.65				
	19.61					13.78				
	18.2					16.23				
	18.45					13.53				
	18.5					11.49				
	18.42					10.29				
	16.3					9.57				
	17.45									
	15.58									
	16.57									
	17.27									
	17.74									
	16.09									

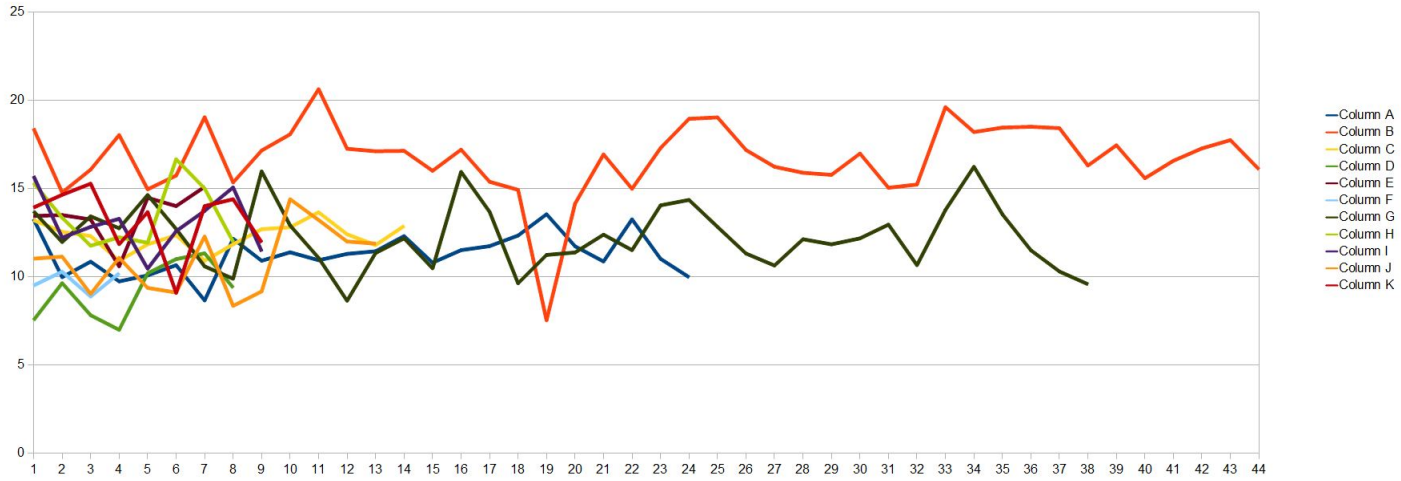

Figure 1.4 : Comma

Question Marks(?)

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
2.59	1.08	1.67	1.88	3.72	1	1.3	1.14	1.9	1.62	1.07
2.17	0.5	1.68	1.45	2.58	0.92	0.99	2.58	0.91	1.44	1.17
2.2	0.27	1.51	1.19	2.64	0.58	0.79	2.95	1	1.44	1.02
2.4	0.96	1.29	1.09	0.47	0.93	1.27	0.69	1.63	1.77	1.35
2.83	1.05	1.66	0.92	0.93		1.38	3.23	1.46	1.34	1.07
2.66	1.54	1.85	1.5	0.9		1.16	3.35	1.87	1.4	3.76
2.58	0.91	1.29	1.1	0.47		0.58	0.69	2.1	1.37	1.09
2.38	1.07	1.66	1.11			0.37	2.58	1.46	0.71	1.35
2.67	0.87	1.35				0.08		0.89	1.39	1.58
2.39	1.07	1.18				0.87			2.47	
2.57	1.31	2.44				0.23			2.21	
2.68	1.45	2.05				0.59			1.86	
2.16	1.21	1.92				0.26			1.46	
2.61	0.58	2.17				0.89				
2.07	1.03					0.16				
2.43	1.02					1.46				
2.48	0.48					0.52				
2.62	0.41					0.47				
2.12	0.64					0.29				
2.22	0.27					0.61				
2.89	1.52					0.43				
2.59	1.49					0.84				
1.84	1.53					1.33				
2.17	0.63					0.81				
	0.07					0.8				
	0.77					1.11				
	0.57					0.88				
	0.6					0.84				
	0.21					0.33				
	1.15					0.62				
	0.16					0.67				
	0.22					0.52				
	1.1					1.07				
	0.86					1.32				
	0.92					1.2				
	0.64					0.37				
	1.28					1.05				
	1.33					0.5				
	0.96									
	0.49									
	0.09									
	0.61									
	1.37									
	1.18									

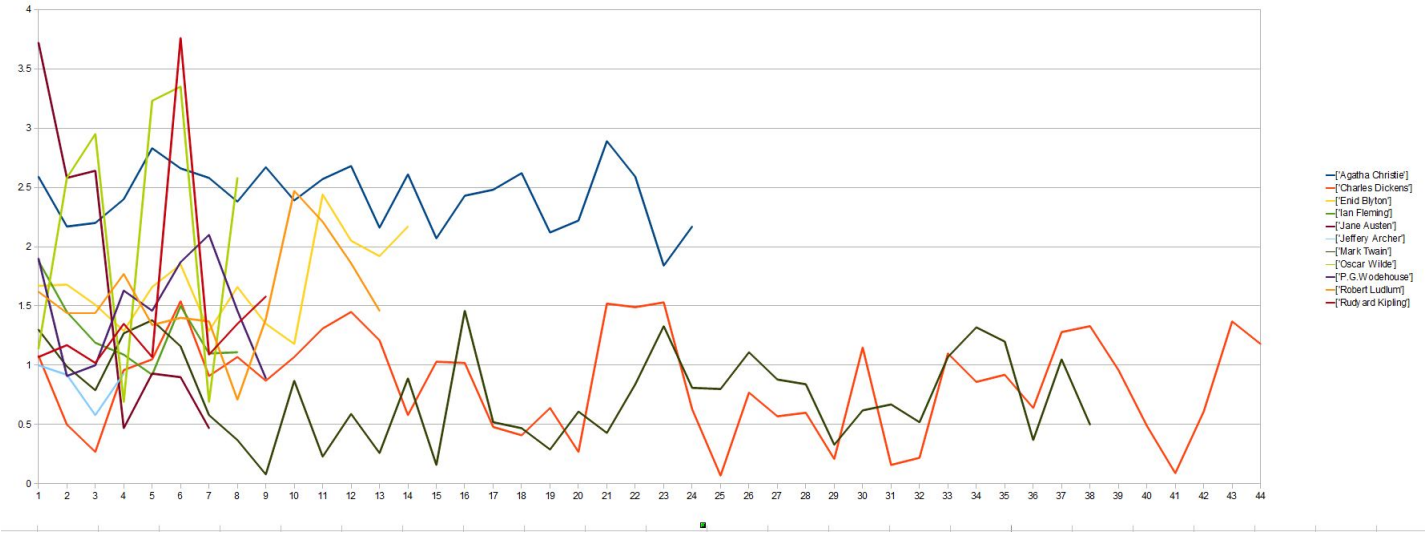

Figure 1.5 : Question-Marks

Opening Quotes(”)

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
13.73	13.34	11.84	0	0	9.48	15.24	0.47	12.29	0	8.32
12.56	5	0	5.26	0	0.08	12.03	0	7.2	0.31	10.42
14.62	0.06	13.39	0	0	5.87	6.76	0	7.85	8.57	6.87
13.74	0.08	10.64	0.01	3.81	9.23	11.65	0.13	10.51	9.11	0
14.83	7.85	13.22	4.4	0		11.29	0	10.14	6.21	0
12.01	3.4	0.08	0.05	0		6.86	0	12.06	6.25	0
12.69	0.11	10.64	0.02	0		4.56	10.42	11.36	6.94	0
10.38	0.03	13.22	0.37			4.4	0	1.75	0.09	10.21
14.79	5.51	0.17				2.73		4.67	7.84	0
13.58	0.1	0.15				9.8			12.44	
14.33	9.3	0				1.68			13.06	
14.77	0.05	0				6.18			9.3	
0.18	0.09	0.01				3.22			7.39	
14.42	0.65	0.07				7.55				
13.3	0.09					3.75				
15.06	0.23					10.78				
0.62	0.5					8.01				
14.46	1.08					5.33				
12.72	7.01					2.53				
15.18	0.22					3.83				
0.43	9.7					5.49				
13.73	8.3					8.96				
12.23	0.23					15.61				
12.56	5.18					6.85				
	0					6.34				
	1.03					6.06				
	0.2					4.58				
	0					11.84				
	0					4.41				
	5.58					5.76				
	1.02					7				
	0.03					6.38				
	0.19					12.19				
	0.25					16.01				
	0.17					15.04				
	3.95					3.75				
	9.38					3.16				
	0.18					4.16				
	0.09									
	0.07									
	2.25									
	4.42									
	0.08									
	8.07									

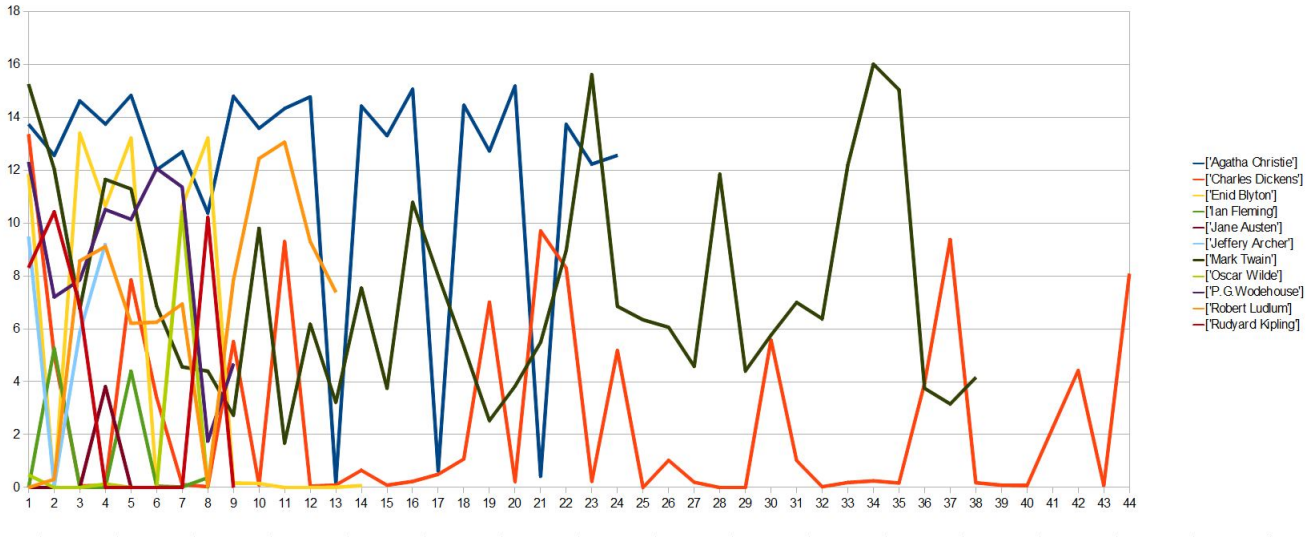

Figure 1.6 : Opening Quote

Hyphens(-)

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
4.51	4.61	2.01	3.25	0.65	0.64	5.99	1.71	1.21	0	4.95
4.23	2.85	2.01	1.23	0.61	1.36	4.22	1.38	3.1	0	4.43
2.9	1.95	2.22	2	0.53	1.44	4.98	1.47	3.35	1.94	3.71
0.72	4.25	1.91	2.51	0.67	0.94	7.69	1.39	3.92	1.15	2.61
3.4	3.33	2.34	1.38	0.5		5.32	3.57	3.02	2.14	3.21
6.12	3.95	4.69	2.92	0.53		3.58	1.33	4.9	2.53	1.8
6.63	2.23	1.91	1.14	0.62		5.26	1.5	3.98	3.33	4.95
6.95	2.66	2.34	2.02			3.02	1.24	0.97	2.23	4.75
0.46	4.5	3.47				4.99		2.33	1.54	2.02
3.54	2.14	3.4				7.17			3.32	
2.91	2.42	3.91				3.81			2.24	
4.47	3.3	4.13				4.1			2.02	
0.51	2.15	4.83				5.15			1.05	
4.4	3.19	4.66				4.45				
2.71	3.64					3.95				
4.42	2.81					4.24				
3	2.04					7.82				
4.41	2.69					4.34				
2.65	2.98					1.72				
1.95	3.23					5.12				
3.89	3.46					5.71				
4.51	3.62					10.65				
7.22	3.07					6.14				
4.23	4.1					7.09				
	2.73					4.1				
	3.98					6.97				
	4.32					3.62				
	3.12					4.92				
	2.8					5.15				
	5.81					4.61				
	2.42					3.77				
	2.55					4.1				
	2.99					5.14				
	2.52					4.11				
	2.69					5.44				
	5.16					4.78				
	3.18					6.36				
	2.4					4.42				
	3.88									
	3.45									
	3.43									
	3.83									
	3.51									
	4.08									

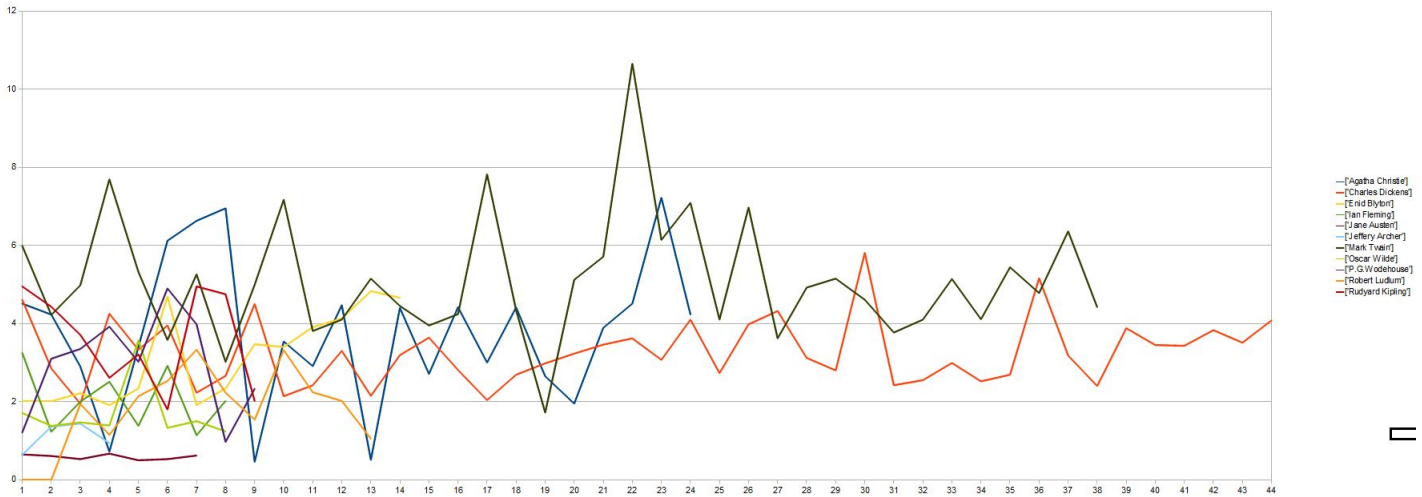

Figure 1.7 : Hyphen

Exclamation Marks(!)

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
1.28	0.79	2.62	0.5	0.78	0	0.8	0.24	0.33	0.72	1.21
1.45	0.35	2.41	0.49	0.49	0.55	1.14	0.86	0.08	1.29	0.59
0.68	0.32	2.66	0.15	0.55	0	1.26	1.83	0.21	0.44	0.52
1.55	1.15	3.38	0.27	0.56	0.03	0.73	0.18	0.27	0.96	1.22
1.64	1.32	2.68	0.62	0.8		0.9	2.66	1.43	0.13	0.99
1.45	1.02	3.27	1.21	1		2.27	1.42	2.71	0.33	3.68
0.83	1.33	3.38	0.84	0.68		0.12	0.72	1.25	0.59	1.49
0.13	1.09	2.68	0.66			0.49	1	0.49	0.5	1.23
0.43	1.02	2.05				1.32		0.22	1.73	1.13
0.54	1.1	1.95				1.32			1.73	
0.27	2.12	3.01				0.27			1.23	
0.46	1.33	2.87				0.55			0.43	
0.22	1.17	4.01				0.18			0.55	
1.18	0.79	3.08				0.57				
0.29	1.1					0.51				
0.69	1.82					1.12				
0.43	0.54					0.88				
1.18	0.3					0.15				
0.12	1.47					0.59				
0.58	0.43					0.42				
1.01	1.29					0.55				
1.28	1.3					1.32				
0.55	1.34					0.82				
1.45	1.29					1.4				
	0.67					0.55				
	1.03					0.1				
	0.9					0.76				
	0.72					0.92				
	0.51					0.68				
	1.32					0.79				
	0.05					0.64				
	0.16					0.55				
	2.07					0.99				
	3.52					0.76				
	3.06					1.17				
	0.86					0.38				
	1.23					0.32				
	1.16					0.24				
	1.02									
	0.56									
	0.64									
	0.78									
	1.53									
	1.06									

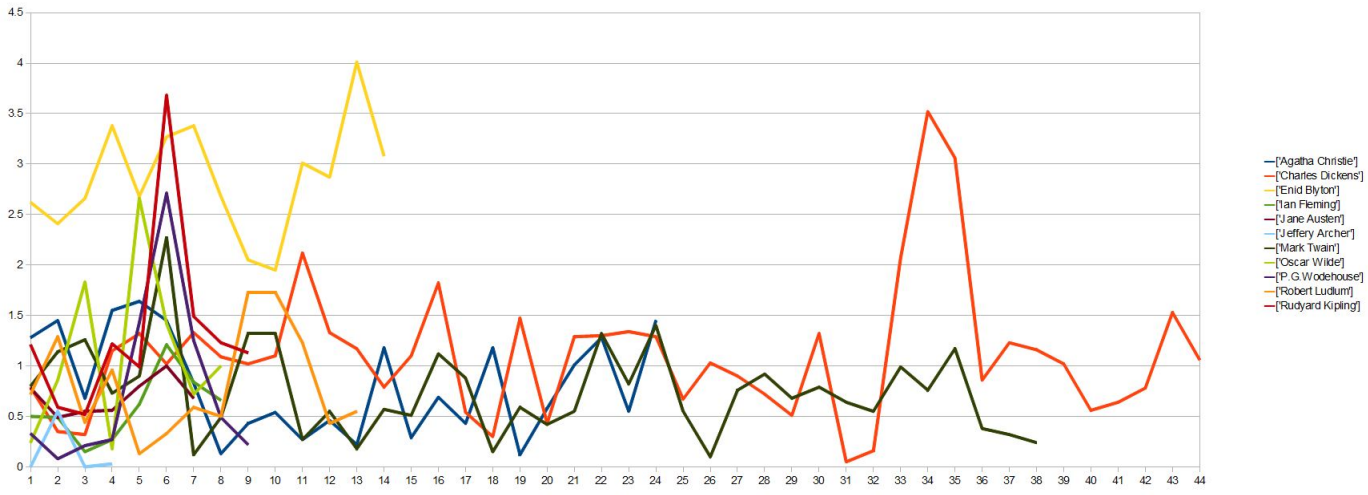

Figure 1.8 : Exclamation Mark

Dots(.)

Agatha Christie	Charles Dickens	Enid Blyton	Ian Fleming	Jane Austen	Jeffery Archer	Mark Twain	Oscar Wilde	P.G.Wodehouse	Robert Ludlum	Rudyard Kipling
14.35	7.74	13.25	14.16	10.07	10.47	8.85	9.49	13.86	14.09	10.18
16.18	6.61	13.61	15.25	9.41	10.3	9.24	16.47	12.62	12.54	11.55
17	5.55	14.12	13.75	9.4	9.12	8.43	16.41	12.79	14.93	9.53
17.59	6.7	13.5	13.6	10.31	10.7	8.72	9.82	13.67	13	11.93
17.84	9.97	14.92	12.39	8.87		8.05	16.27	13.31	15.02	9.45
15.93	7.81	14.62	12.81	7.39		7.96	9.07	12.81	16.7	30.69
17.12	6.8	13.5	11.93	6.86		8.55	10.49	14.78	11.73	8.25
16.89	7.88	14.92	11.93			7.54	15.11	14.28	17.11	11.07
17.41	8.45	12.42				6.31		11.55	13.64	11.88
16.88	10.11	12.27				9.1			11.47	
17.51	8.75	15.79				10.7			12.78	
17.45	11.18	15.6				8.47			14.46	
17.38	10.98	14.1				8.76			13.03	
16.39	7.85	13.96				8.45				
17.17	7.64					7.5				
17.87	7.64					8.51				
14.06	6.22					7.79				
16.38	5.76					7.28				
16.18	6.1					7.71				
16.57	6.63					6.94				
16.68	8.93					7.96				
14.35	10.22					8.36				
18.99	8.52					9.07				
16.18	8.4					6.89				
	5					8.59				
	6.71					7.87				
	7.02					7.79				
	7.04					7.86				
	4.53					7.52				
	7.55					7.6				
	4.64					8.69				
	4.43					9.57				
	9.46					7.9				
	8.98					8.57				
	8					8.65				
	6.47					7.6				
	8.17					8.64				
	9.92					8.84				
	6.56									
	5.94									
	7									
	6.71									
	10.28									
	7.86									

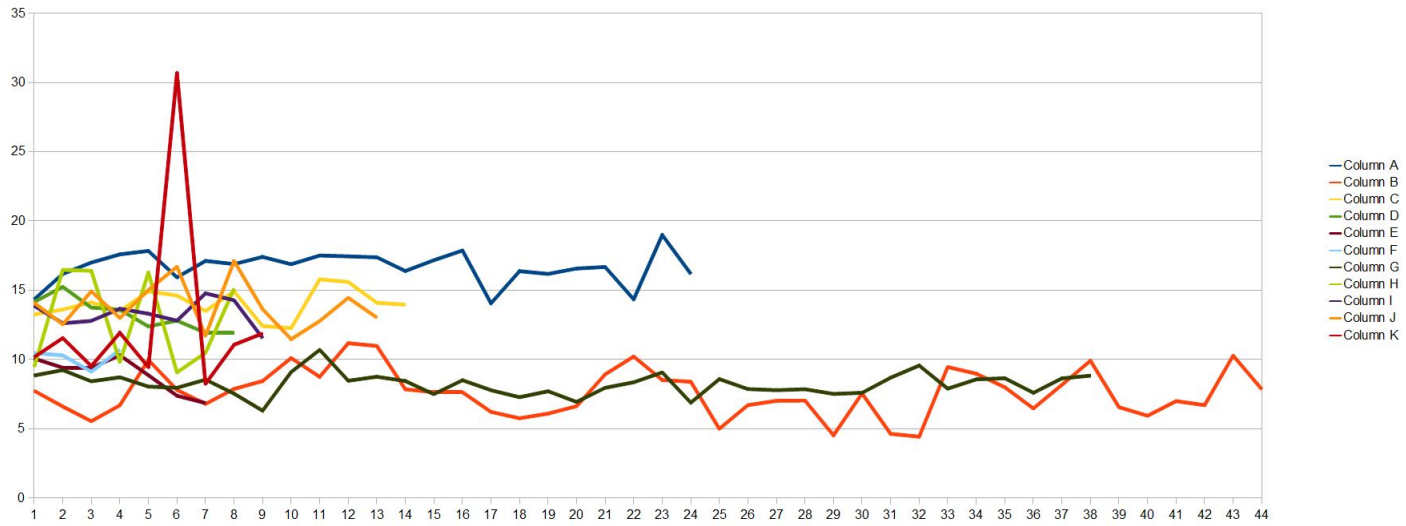

Figure 1.9 : Dots

FORMULA USED

In our final algortihm we multiplied all the relevant data frequencies after scaling them by a suitable factor and the two authors whose scores were highest among all the authors were reported as the best possbile guess as the actual author for the test case novel.

We rate the authors based on the product of relative frequencies using the formula:

R(Author)=R(Genre)*R(Era)*R(Depth)*R(Length)*R(Dots)*R(QuestionMarks)

where:

R(Author) = Relative ratio of an author. This is what is calculated for each author and compared to predict the authors.

R(Genre) = Total number of books of an author of the same genre as the test novel/Total number of novels of the author in the dataset

R(Era) =Total number of books of an author of the same era as the test novel/Total number of novels of the author in the dataset

R(Depth) =Total number of books of an author having depth in range of ± para-depth (=0.6 by analysis of Figure 1.2) of the average depth of the newnovel/ Total number of novels of the author in the dataset

R(Length) =Total number of books of an author having Length in range of ± para-length (=2.0 by analysis of Graph 1.1) of the average length of the newnovel/Total number of novels of the author in the dataset

R(Dots) =Total number of books of an author having Length in range of ± para-dots (=2.0 by analysis of Graph 1.9) of the relative number of dots in the newnovel/Total number of novels of the author in the dataset

R(QuestionMarks) =Total number of books of an author having Length in range of ± para-questionmarks (=0.5 by analysis of Graph 1.5) of the relative number of question marks of the newnovel/Total number of novels of the author in the dataset

If the author is using relative number of hyphens =0.8 then we increase the weightage of the author being Jane Austen(as evident from Figure 1.7) by a factor of 10.

Similarly if the author is using a relative number of colons =0 then we increase the weightage of the author being Enid Blyton(as evident from Figure 1.3) by a factor of 10.

Using the above formula, the 2 authors with highest product come out to be our predictions.

EXAMPLE 1

The file The Adventurous Four Again is of Era:20th century

The file The Adventurous Four Again is Romance
The file The Adventurous Four Again Maybe crime also
The file The Adventurous Four Again Maybe Fantasy also

Punctuation marks used in the text are:
[[',', 10.92], ['.', 13.95], ['"', 6.67], ['"', 11.67], [':', 0.01], [';', 0.01], ['?', 1.52], ['/', 0.0], ['*', 0.01], ['!', 3.46], ['(', 0.0], ['‘', 0.0], ['-', 1.56], ['@', 0.0], ['[', 0.0]]

Average depth of a sentence is 8.09984399376
All values below are for the 11 authors in the order Agatha Christie,Charles Dickens,Enid Blyton,Ian Fleming,Jane Austen,Jeffery Archer,Mark Twain,Oscar Wilde,P.G.Wodehouse,Robert Ludlum and Rudyard Kipling respectively.

R(ERA)
0.958333333333
0.522727272727
1.0
1.0
0.285714285714
0.75
0.868421052632
0.625
0.777777777778
1.0
0.888888888889

R(GENRES)
1.16666666667
0.704545454545
1.28571428571
0.125
1.0
1.0
0.657894736842
1.0
1.11111111111
1.0
1.0

R(SENTENCE LENGTH)
0.75
0.0
0.928571428571
0.375
0.142857142857
0.0
0.0
0.5
0.666666666667
0.846153846154
0.222222222222

R(SENTENCE DEPTH)
0.916666666667
0.0
0.857142857143
0.625
0.0
0.0
0.0
0.5
0.333333333333
0.769230769231
0.111111111111

R(DOTS)
0.166666666667
0.0
1.0
0.75
0.0
0.0
0.0
0.125
0.888888888889
0.692307692308
0.0

R(QUESTION MARK)
0.0416666666667
0.409090909091
0.785714285714
0.875
0.0
0.0
0.289473684211
0.125
0.555555555556
0.769230769231
0.777777777778

Multiplied array containing each author's relative ratios:
[5.337938850308642, 0.0, 8040.399833402749, 19.22607421875, 0.0, 0.0, 0.0, 2.44140625, 94.8364917272096, 346.62651867931794, 0.0]

Above array in sorted order:
[0.0, 0.0, 0.0, 0.0, 0.0, 2.44140625, 5.337938850308642, 19.22607421875, 94.8364917272096, 346.62651867931794, 8040.399833402749]

Final Prediction:
1. Enid Blyton
2. Robert Ludlum

The first prediction is correct in this example.

EXAMPLE 2

The file Adventurous Four is of Era:20th century
The file Adventurous Four is Crime

Punctuation marks used in the text are:
[[',', 9.89], ['.', 12.86], ['"', 5.15], ['"', 11.17], [':', 0.03], [';', 0.02], ['?', 1.12], ['/', 0.0], ['*', 0.0], ['!', 3.12], ['(', 0.0], ['‘', 0.0], ['-', 5.77], ['@', 0.0], ['[', 0.0]]

Average depth of a sentence is 8.58781362007
All values below are for the 11 authors in the order Agatha Christie,Charles Dickens,Enid Blyton,Ian Fleming,Jane Austen,Jeffery Archer,Mark Twain,Oscar Wilde,P.G.Wodehouse,Robert Ludlum and Rudyard Kipling respectively.

R(ERA)
0.958333333333
0.522727272727
1.0
1.0
0.285714285714
0.75
0.868421052632
0.625
0.777777777778
1.0
0.888888888889

R(GENRES)
0.5
0.136363636364
0.0714285714286
0.125
0.0
1.0
0.368421052632
0.0
0.111111111111
1.0
0.444444444444

R(SENTENCE LENGTH)
0.333333333333
0.0227272727273
0.928571428571
0.75
0.142857142857
0.0
0.0
0.375
0.777777777778
1.0
0.444444444444

R(SENTENCE DEPTH)
0.916666666667
0.136363636364
1.0
1.0
0.0
0.0
0.0
0.25
0.666666666667
0.846153846154
0.444444444444

R(DOTS)
0.125
0.0454545454545
0.714285714286
0.875
0.0
0.0
0.0
0.0
1.0
0.692307692308
0.444444444444

R(QUESTION MARK)
0.0
0.659090909091
0.357142857143
0.875
0.285714285714
0.75
0.552631578947
0.375
0.555555555556
0.615384615385
0.888888888889

Multiplied array containing each author's relative ratios:
0.0, 0.006618236332251612, 169.2003331945023, 71.77734375, 0.0, 0.0, 0.0, 0.0, 24.89457907839252, 360.49157942649066, 30.829386517035747]

Above array in sorted order:
[0.0, 0.0, 0.0, 0.0, 0.0, 0.006618236332251612, 24.89457907839252, 30.829386517035747, 71.77734375, 169.2003331945023, 360.49157942649066]

Final Prediction:
2. Enid Blyton
1. Robert Ludlum

In this example, the second prediction was correct.

EXAMPLE 3

The file The Altman Code is of Era:18th century
The file The Altman Code is Crime

Punctuation marks used in the text are:
[[',', 11.4], ['.', 15.96], ['"', 4.7], ['"', 9.26], [':', 0.14], [';', 0.02], ['?', 1.65], ['/', 0.0], ['*', 0.0], ['!', 0.23], ['(', 0.01], ['‘', 0.0], ['-', 2.12], ['@', 0.0], ['[', 0.0]]

Average depth of a sentence is 8.21016949153
All values below are for the 11 authors in the order Agatha Christie,Charles Dickens,Enid Blyton,Ian Fleming,Jane Austen,Jeffery Archer,Mark Twain,Oscar Wilde,P.G.Wodehouse,Robert Ludlum and Rudyard Kipling respectively.

R(ERA)
0.0416666666667
0.477272727273
0.0
0.0
0.714285714286
0.25
0.105263157895
0.125
0.222222222222
0.0
0.0

R(GENRES)
0.5
0.136363636364
0.0714285714286
0.125
0.0
1.0
0.368421052632
0.0
0.111111111111
1.0
0.444444444444

R(SENTENCE LENGTH)
0.958333333333
0.0
0.857142857143
0.125
0.0
0.0
0.0
0.5
0.444444444444
0.461538461538
0.222222222222

R(SENTENCE DEPTH)
0.958333333333
0.0681818181818
0.857142857143
0.875
0.0
0.0
0.0
0.5
0.444444444444
0.769230769231
0.222222222222

R(DOTS)
0.958333333333
0.0
0.5
0.25
0.0
0.0
0.0
0.5
0.222222222222
0.461538461538
0.0

R(QUESTION MARK)
0.125
0.25
0.857142857143
0.5
0.0
0.0
0.210526315789
0.0
0.666666666667
0.769230769231
0.444444444444

Multiplied array containing each author's relative ratios:
[2.292020821277006, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.7225637464930253, 0.0, 0.0]

Above array in sorted order:
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.7225637464930253, 2.292020821277006]

Final Prediction:
1. Agatha Christie
2. P.G.Wodehouse

In this example, the original author was Robert Ludlum, but both the predictions came out to be wrong. It is notable that apart from era, in all other terms Robert Ludlum had considerable values.

CODING DETAILS

Most of the coding for our project was done on python. We chose Python as the coding platform for our project because of the relative ease in handling arrays in python and also because of a large plethora of libraries of python available at our disposal.
These features of Python make it an ideal language for **Natural Language Processing**(NLP). The Stanford Parser used by us in the project was developed in Java and we therefore had to use Jython as a linking bridge between the two languages. The data of the parsed sentences was used by the python code.

Our program is divided into the following python code files:

- **stanfordparser.py** : To call the jython script myscript.py.
- **myscript.py** : To start the parsing of the sentences of the file using the stanfordparser.jar(Note: We used the englishPCFG parser of the stanford parser because of it's high speed as it was one of our main concerns.)
- **readdata.py** : To form an array of the previously available data we had. The array formed was our dataset.
- **punctuation.py**: To report the relative use of the punctuations in the novels. It returned an array comprising of all the punctuations and their relative usage in the program.
- **newnovels.py** : This program prints the percentage of correct classifications of the new test case files of the authors.
- **newnovel.py** : To be called by newnovels.py to collect data on new indivisual
- **genres.py**: Calls genrecheck.py to find the genre of the novel
- **finaldata.py**: Contains the final dataset (array returned by readdata.py).
- **avglength.py**: Returns the average sentence length of sentences used by the author in a particular novel.
- **era.py**:Returns the era of the novels.

RESULTS

S.No.	Novel Name	Predicted Author 1	Predicted Author 2	Actual Author	Correct/Incorrect
1	casino royale	Ian Fleming	P.G.Wodehouse	Ian Fleming	✓
2	doctor no	Ian Fleming	Robert Ludlum	Ian Fleming	✓
3	for your eyes only	Ian Fleming	Robert Ludlum	Ian Fleming	✓
4	from russia with love	Ian Fleming	Robert Ludlum	Ian Fleming	✓
5	goldfinder	Ian Fleming	Robert Ludlum	Ian Fleming	✓
6	live and let die	Ian Fleming	Robert Ludlum	Ian Fleming	✓
7	Moonraker	Ian Fleming	P.G.Wodehouse	Ian Fleming	✓
8	Covert One	Robert Ludlum	P.G.Wodehouse	Robert Ludlum	✓
9	The Bourne Legacy	Robert Ludlum	P.G.Wodehouse	Robert Ludlum	✓
10	The Altman Code	Agatha Christie	P.G.Wodehosue	Robert Ludlum	X
11	Covert One-The Moscow Vector	Robert Ludlum	Enid Blyton	Robert Ludlum	✓
12	The Matarese Countdown	Robert Ludlum	Agatha Christie	Robert Ludlum	✓
13	Trevayne	Robert Ludlum	Agatha Christie	Robert Ludlum	✓
14	The Parcifal Mosaic	Robert Ludlum	Ian Fleming	Robert Ludlum	✓
15	The Bourne Ultimatum	Robert Ludlum	Agatha Christie	Robert Ludlum	✓
16	My Man Jeeves	P.G.Wodehouse	Enid Blyton	P.G.Wodehouse	✓
17	Piccadily Jim	P.G.Wodehouse	Robert Ludlum	P.G.Wodehouse	✓
18	Six Short Stories	P.G.Wodehouse	Jeffery Archer	P.G.Wodehouse	✓
19	Something Fresh	P.G.Wodehouse	Rudyard Kipling	P.G.Wodehouse	✓
20	Stiff Upper Lip Jeeves	P.G.Wodehouse	Rudyard Kipling	P.G.Wodehouse	✓
21	Gold Bat	P.G.Wodehouse	Rudyard Kipling	P.G.Wodehouse	✓
22	The Spring Suit	Ian Fleming	Robert Ludlum	P.G.Wodehosue	X
23	Adventurous Four	Robert Ludlum	Enid Blyton	Enid Blyton	✠
24	The Adventurous Four Again	Enid Blyton	Robert Ludlum	Enid Blyton	✓
25	Valley of Adventure	Enid Blyton	P.G.Wodehouse	Enid Blyton	✓
26	Barney Junior Mystery	Enid Blyton	Agatha Christie	Enid Blyton	✓
27	Mystery of Burnt Cottage	Agatha Christie	Enid Blyton	Enid Blyton	✠
28	The Mystery of Disapperaing Cat	Enid Blyton	P.G.Wodehouse	Enid Blyton	✓
29	The Mystery of Holly Lane	Enid Blyton	Agatha Christie	Enid Blyton	✓
30	The Mystery of Banshee Towers	Enid Blyton	P.G.Wodehouse	Enid Blyton	✓
31	The Twins At St. Clare's	Enid Blyton	P.G.Wodehouse	Enid Blyton	✓
32	Summer Term at St. Clare's	Enid Blyton	P.G.Wodehouse	Enid Blyton	✓

S.No.	Novel Name	Predicted Author 1	Predicted Author 2	Actual Author	Correct/Incorrect
33	The Second Form At St. Clare's	Enid Blyton	P.G.Wodehouse	Enid Blyton	✓
34	Mystery of the Secret Room	Enid Blyton	Robert Ludlum	Enid Blyton	✓
35	Mystery of Spiteful Letters	Enid Blyton	Agatha Christie	Enid Blyton	✓
36	Mystery of Hidden House	Enid Blyton	Agatha Christie	Enid Blyton	✓
37	A Holiday Mystery	Agatha Christie	Robert Ludlum	Agatha Christie	✓
38	Evil Under the Sun	Agatha Christie	Robert Ludlum	Agatha Christie	✓
39	Five Little Pigs	Enid Blyton	Robert Ludlum	Agatha Christie	X
40	Lord Edgware	Agatha Christie	Robert Ludlum	Agatha Christie	✓
41	Murder On the Links	Agatha Christie	Robert Ludlum	Agatha Christie	✓
42	Murder in Mesopotamia	Ian Fleming	Robert Ludlum	Agatha Christie	X

Novels shown with ✠ are those in which the second prediction was correct.

We tested the program on new novels (total 42 in number) by these authors (novels that were not previously used in our dataset of 198 novels.) Some of the observations for some of the authors are presented here:

- **Enid Blyton**: We tested on 14 novels of Enid Blyton and the program came up with a 100% accuracy.
- **Agatha Christie**: Out of the 6 new novels of Agatha Christie we tested we came up with 4 correctly classified.
- **Ian Fleming**: Out of the 7 novels of Ian Fleming all of them were correctly classified as being written by Ian Fleming giving a 100% accuracy in this case.
- **P.G. Wodehouse**: Our of the 7 test cases 6 were correctly classified as being written by P.G.Wodehouse thereby acheving an accuracy of 85.7
- **Robert Ludlum**: Out of the 8 new novels of Robert Ludlum(whose data was not included in the dataset previously)7 were correctly classified as being written by Robert Ludlum and only one was classified in correctly.

The overall accuracy of our program on the test data was 90.48% (when including the second author predicted) and 85.71% (when taking only the first prediction).

FAILURES

There were some other parameters also, that we tried but were not good enough either because they were not consistent for all the works of the authors themselves or they were not able to successfully differentiate between the differnt authors. Some of those parameters are listed here:

- **Average word length**: This means the sum of all the lengths of the words used in the novel/total number of the words in the novel. This was not a good parameter. Although they were quite consistent among all the different works of the author but they were also nearly same for many authors nearly tending to 4.5.
- **Some punctuations like**:opening quotes and exclamation mark as depicted by the graph 1.6 and 1.8 were not consistent among different works of the author and were also not successfully differentiating between the different authors.
- **Bigrams and Trigrams**: The authors were not using any fixed or particular bigrams or trigrams with consistency in their works. This idea had to be dropped alltogether.

FURTHER SCOPE AND LIMITATIONS

The dataset of authors can be increased but since we are using only a very limited number of parameters; if this project is scaled to a very large number of authors the results might be conflicing. However, if some new parameters can be identified which satisfy both the properties:

- Consistency for all the works of the author
- Ability to distinguish between the different authors

Then the number of authors can be increased to a large extent.

The main limitation one has to face while working with classification of authors is that one has to grasp the writing style of the author rather than just judge them by the type of vocabulary used. For this one will have to explore the semantic as well as syntactic components of their works.

Another limitation is that the system should be dynamic ie. it should update itself whenever a test novel is given. We on the other hand have a static system so far.

Also finding good parameters can be a very daunting task. We define a good parameter which satisy both the aforementioned properties.

APPENDIX

DATASET USED

Agatha Christie

- The Mysterious Affair at Styles
- The Secret Adversary
- The Murder at the Vicarage
- Appointment with Death
- hercules poirots christmas
- One Two Buckle My Shoe
- evil under the sun
- DEATH COMES AT THE END
- A Caribbean Mystery
- 4.50 From Paddington
- A Pocket Full Of Rye
- At Bertram's Hotel
- By the Pricking of My Thumbs
- Christie, Agatha easy to kill
- Christie, AGATHA endless night
- Death In The Air
- early cases of herquells poirot
- Easy To Kill
- Elephants Can Remember-poirot
- Partners in Crime
- the case book of herquelles poirot
- The Mysterious Affair at Styles
- The Mysterious Mr Quin
- The Secret Adversary

Charles Dickens

- A Message From the Sea
- All the Year Round
- American Notes
- Barnaby Rudge
- Bleak House
- Doctor Marigold
- Dombey and Son
- George Silverman's Explanation
- Going Into Society
- Hard Times
- Haunted Man Ghost's Bargain
- Holiday Romance
- Hunted Down
- Lazy Tour of Two Idle Apprentices

- Little Dorrit
- Martin Chuzzlewit
- Master Humphrey's Clock
- Miscellaneous Papers
- Mrs. Lirriper's Legacy
- Mudfog & Other Sketches, by Charles Dickens
- Mugby Junction, by Charles Dickens
- No Thoroughfare, by Dickens & Collins
- Our Mutual Friend, by Charles Dickens
- Perils of Certain English Prisoners, by Charles Dickens
- Pictures From Italy, by Charles Dickens
- Reprinted Pieces, by Charles Dickens
- Sketches by Boz, by Charles Dickens
- Sketches of Young Couples, by Charles Dickens
- Sketches of Young Gentlemen, by Charles Dickens
- Somebody's Luggage, by Charles Dickens
- Speeches Literary and Social, by Charles Dickens
- Sunday Under Three Heads, by Charles Dickens
- The Battle of Life, by Charles Dickens
- The Chimes, by Charles Dickens
- The Cricket on the Hearth, by Charles Dickens
- The Holly-Tree, by Charles Dickens
- The Lamplighter, by Charles Dickens
- The Mystery of Edwin Drood, by Charles Dickens
- The Old Curiosity Shop, by Charles Dickens
- The Uncommercial Traveller
- The Wreck of the Golden Mary
- thesevenpoortravellers
- To Be Read at Dusk
- Tom Tiddler's Ground

Enid Blyton

- Mystery of the Missing Necklace
- The Circus of Adventure
- Adv 04 - Sea of Adventure
- Adv 05 - Mountain of Adventure
- Adv 06 - Ship of Adventure
- Adv 07 - Circus of Adventure
- Adventure 05 - The Mountain of Adventure
- Adventure 06 - The Ship of Adventure
- St Clare's 05 - Claudine at St Clare's (b)
- St Clare's 06 - Fifth Formers at St Clare's
- Mystery 09 - Mystery of the Vanished Prince
- Mystery 10 - Mystery of the Strange Bundle
- Mystery 12 - Mystery of Tally-Ho Cottage
- Mystery 13 - Mystery of the Missing Man

Ian Fleming

- hildebrand rarity
- the man with golden gun
- diamonds are forever
- for your eyes only james bond
- octopussy
- on her majesty's secret service
- the spy who lovedme
- you only live twice

Jane Austen

- emma
- mansfield park
- pride and prejudice
- sandition
- sense and sensibility
- northanger abbey
- persuasion

Jeffery Archer

- as the crow flies
- kane and abel
- quiver full of arrows
- the fourth estate

Mark Twain

- Adventures of Huckleberry Finn
- Adventures of Tom Sawyer
- Alonzo Fitz And Other Stories
- American Claimant
- Captain Stormfield's Visit To Heaven
- Carnival of Crime in CT
- Christian Science
- Curious Republic Of Gondour And Sketches
- Dog's Tale
- Double Barrelled Detective Story
- Extracts From Adam's Diary
- Fenimore Cooper's Literary Offences
- Following The Equator
- Gilded Age - A tale of today
- Goldsmith's Friend Abroad Again
- Horse's Tale
- How To Tell A Story And Other Essays
- In Defence Of Harriet Shelley
- Innocents Abroad

- Is Shakespeare Dead - From my autobiography
- Life on the Mississippi
- Man That Corrupted Hadleyburg
- Mark Twain - Adventures of Huckleberry Finn
- Mark Twain - The Prince and the Pauper
- Mysterious Stranger
- On The Decay Of The Art Of Lying
- Personal Recollections Of Joan Of Arc vol 2
- Pudd'n'head Wilson
- Roughing It
- Sketches New And Old
- Some Rambling Notes Of An Idle Excursion
- Stolen White Elephant
- Those Extraordinary Twins
- Tom Sawyer, Abroad
- Tom Sawyer, Detective
- Tramp Abroad
- What Is Man - and Other Essays
- What Paul Bourget Thinks Of Us

Oscar Wilde

- A House of Pomegranates
- A Woman of No Importance
- An Ideal Husband
- Intentions
- Lady Windermere's Fan
- The Duchess of Padua
- The Happy Prince and Other Tales
- The importance of being earnest

P.G.Wodehouse

- love me love my dog
- plumpunch
- shortstories
- thefifteenthman
- damsel in distress
- indiscretion of archie
- intrusion of jimmy
- ladiesandgentlemenlayers
- signsandportents

Robert Ludlum

- identity
- supremacy
- Matarese Dynasty 01 - The Matarese Circle
- The Aquitaine Progression
- The Holcroft Covenant
- The Matlock Paper
- The Prometheus Deception
- The Rhineman Exchange
- The Road To Gandolfo
- The Road to Omaha
- The Scorpio Illusion
- The Sigma Protocol
- The Tristan Betrayal

Rudyard Kipling

- junglebook
- The Man Who Would Be King
- actions and reactions
- days-work
- gadsby
- justsostories
- Captain Courageous
- light-failed

REFERENCES

[1] Marcia Fissette (2010) : Author Identification in Short Texts, `www.ru.nl/publish/pages/641151/fissette_m_bathesis10.pdf`

[2] C.Chaski. Who wrote it? steps towards a science of authorship identification. *National Institute of Justice Journal*, September 1997.

[3] M. Corney, A. Anderson, G. Mohay, and O. de Vel. Identifying the authors of suspect mail. *Communications of the ACM*, 2001.